Assessing Test Reliability:

Comparing Two Versions of Reading Comprehension Test in the TOEFL test

Zhang Heng
Kristianstad University
The School of Teacher Education
English IV, spring 2010
D-essay in English Didactics
Tutor: Jane Mattisson

Table of Contents

1. Introduction

ETS, Educational Testing Service, is a non—benefit making US company founded in 1947. According to its official website, ETS has "developed, administered and scored more than 50 million tests annually… in more than 180 countries and at over 9,000 locations worldwide."[1] Aiming at measuring and improving knowledge and skills in English language, ETS prides itself on offering fair and valid tests. One of the most prestigious ETS tests is the TOEFL test.

TOEFL, abbreviated from Test of English as a Foreign Language, has been adopted as English proficiency test by more than 7,300 univeristies and colleges in around 130 countries, including the U.S.A, UK, Canada, Australia, Germany, and Holland, and the TOEFL score has become a crucial qualification for college entrance applications. TOEFL, therefore, has a large number of test takers each year all over the world. An official statistic issued recently shows that by 2009, over 2.4 million students from all nationalities have entered for the TOEFL test and the figure will increase much faster in future years.[2]

As a TOEFL test can affect more and more test takers' future and plays such an important role in college entrance applications, studies on the TOEFL test are useful and necessary. First of all, the design principle of TOEFL should be made clear. There are two co-existing versions of TOEFL: the internet-based test (abbreviated IBT) and the paper-based version (abbreviated PBT). And one of them, the IBT is described in *Bulletin for IBT (internet-based test) of TOEFL in 2010-2011* to be of academic purpose:

> The TOEFL test measures the ability of non-English speakers to communicate in English *in an academic setting*. It accurately measures how well students can read, listen, speak, and write in English in the college or university *classroom*.

"In an academic setting" and "classroom" are italicized in order to emphasize that the test takers' purposes make the TOEFL test to be academic, testing the potential abilities

[1] To learn more about ETS, access ETS official website: www.ets.org

[2] These figures are officially issued by the official TOEFL website: http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect_ets.org_toefl [Accessed 25th April, 2010].

in learning subjects in English colleges. Accordingly, the four parts in the TOEFL test, i.e. reading, listening, speaking, and writing, therefore, are all designed to test whether test takers tend to be able to use English in academic environment.

The reading test in the IBT, for instance, "measures the ability to *understand academic* reading material" (*Bulletin for IBT of TOEFL in 2010-2011,* 2009: 3). There are two key words in this principle: "understand" and "academic", which are italicized by me for emphasis. The latter one requires that the topics of reading materials should be related to scientific or technological subjects; the former one state, in dealing with reading materials, what aspect is tested: test takers should "understand" the reading materials, not talk about or memorize the reading materials.

For the PBT, on the other hand, there are no explicit descriptions for its design purpose. However, the following quotation indicates that the PBT shares a similar test purpose with the IBT:

> ETS offers the Internet-based version (TOEFL IBT™) in most locations. The paper-based version (TOEFL PBT) will continue to be offered to supplement the TOEFL IBT test center network. (*Bulletin for PBT of TOEFL in 2010-2011,* 2009: 3)

The quotation demonstrates the fact that in areas where internet access is either limited or non-existent, the PBT is regarded as an adequate and equal substitute. Additionally, this quotation implies it is assumed by the TOEFL organization that the IBT and the PBT are comparable in that a candidate will receive almost exactly the same score irrespective of the type of test taken. The tone of the quotation suggests that TOEFL wants to offer the IBT version and only sees the PBT as an addition in case of need.

Since TOEFL scores are more and more widely-adopted by colleges in making decisions, the co-existence of the IBT and the PBT in the name of the TOEFL test makes the analysis of the consistency of their results a necessity. The three influential factors of the test result i.e. test design feature, test difficulty and scoring system will be analyzed comparatively between IBT and PBT in assessing the TOEFL test reliability.

1.1 Aim

This paper analyzes the two test forms used by TOEFL: IBT and PBT. The analysis will focus on the reading comprehension section, its design features, content, and scoring results. The aim is to assess the reliability of the two test forms as well as to identify factors influencing candidate performance in the reading comprehension test. Three factors are identified: test setting, test difficulty and scoring methods and results. The latter two will be focused on because test difficulty consistency directly decides the test result consistency. And as the goal of the candidate is to achieve as high a score as possible, and success is measured in terms of numbers, score reliability is a primary concern for both candidate and examining body alike.

1.2 Material

Test descriptions of reading tests in TOEFL IBT and PBT are provided in Bulletins of the IBT and the PBT, which are found and downloaded freely from the official TOEFL website. The website also provides **sample tests** of reading comprehension for both TOEFL IBT and PBT. For each of the two versions, there is only one copy of sample reading test, including one passage, 10 to 14 questions, and key answers for the questions.

Scoring reports of TOEFL IBT and PBT in recent five years are also electronically published on the official TOEFL website. Statistics of scores for the IBT and the PBT for 2009, the latest, are selected to analyze in this essay.

Electronically published on the official TOEFL website, *TOEFL Monograph Series* provide academic studies on the TOEFL test from various aspects, including TOEFL reading test studies. This magazine was consulted during the composition process of this essay.

1.3 Method

To begin with, there will be an exact comparison between the IBT and the PBT of the TOEFL test in **test settings**, including physical settings, test delivering modes, instructions, structures and time allotments. At the end of this part, there might be a conclusion on which of the two types of test settings tends to benefit to the TOEFL test takers of either the IBT or the PBT.

In step two, there will be a comparison between reading **test difficulty** in the IBT and the PBT. Factors that can affect reading test difficulty will be examined, including text and question difficulties, text length, and relationship between question and text. At the end of this part, it is possible to find out that whether reading tests in the two versions are of the same difficulty levels and if not, which one is probably more difficult.

Step three will follow step two consistently. The TOEFL **scoring reports**, showing the scores achieved by the TOEFL test takers all over the world in 2009, will be analyzed in this step. Scoring reports for the TOEFL test takers of different genders, different nationalities and different educational backgrounds will be comparatively studied on, between the IBT and the PBT average scores. At the end of this part, the TOEFL test score reliability will be discussed in order to explicate whether the IBT and the PBT scores measure test takers' reading comprehension abilities consistently.

Finally, in comparing results from step two and step three, there will be a discussion on the reliability[3] of the TOEFL reading test within the two versions. If the results of scoring reports analysis are correspondent to test difficulty and test setting analysis results, the TOEFL test as a whole then tends to rate and judge test takers' language abilities consistently in the two versions. Therefore, the TOEFL test can be considered as a reliable one. If there is inconsistency between result of scoring analysis and result of test setting and test difficulty analysis, possible influential factors will be discussed subsequently.

In this essay, no practical tests will be done and no test takers will be involved under observation. The analysis is based on documents that are published by the official TOEFL website.

[3] To know exact definition of 'reliability', see Section 2.1.1.

2. Theoretical background

As two different versions co-exist and the TOEFL test influences more and more people's future, it is necessary to assess whether the TOEFL test with two versions is reliable, because to ensure the consistency and fairness of the TOEFL test results matters not only the TOEFL test itself but also every test taker's life. This section firstly provides theories concerning test reliability in general. Then in the following, there are specific theories concerning the reading comprehension test reliability. Factors affect the reading test reliability are enlisted and explicated in this section, providing possible ways for the reading test reliability assessment.

2.1 Reliability and test usefulness

Test usefulness defines what a test intends to do by its design and development. There are six aspects of test usefulness: reliability, validity, authenticity, interactiveness, impact and practicality.

Usefulness=**Reliability** + Construct validity + Authenticity + Interactiveness + Impact + Practicality

Table 1 A graphic representation of test usefulness from Bachman & Palmer (1996:18)

In the table, which is quoted from Bachman and Palmer (1996: 17), reliability is enlisted as the first one among all the other aspects. It implicates that reliability is a crucial aspect in test usefulness and is worth to do research on.

2.1.1 Test reliability

Test reliability refers to the consistency of score and rank order of test takers' from a test when it is administered in different situations (Bachman and Palmer, 1996: 19-20). It means that if a test taker achieves quite similar scores across different characteristics

of testing situations[4], the test is considered a reliable one. The following table also shows what reliability means:

Scores on test tasks with characteristics X
↕ Reliability
Scores on test tasks with characteristics X'

Table 2 Test reliability

The two sets of test task characteristics (X and X') show that test X and test X' are correspondent in content but are different in incidental ways. If the rank orders of test takers in test X and test X' are the same, we say that test X is reliable. Hughes (2003: 36) states the following to explain reliability. Interpreting the table above, Hughes explains reliability as follows:

> What we have to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more reliable the test is said to be.

It is to say that the extent of reliability depends on the consistency of the rank and scores they achieve at different times in different situations from the same language test. Take the TOEFL test to illustrate: the PBT takes place six times each year and the IBT takes place for more than thirty times per year all over the world. When the same test taker takes randomly two TOEFL tests in a short period of time, and if the results achieved from the two tests implicate similarly the test taker's English ability, the TOEFL test is then considered reliable.

As aforesaid, reliability is an essential quality of test scores, and as test design is more controllable by the test designer in comparing test takers' performance and their topical knowledge, researchers as Bachman and Palmer argue that "in designing and developing language tests, we [should] try to minimize variations in the test task characteristics that do not correspond to variations in TLU[5] tasks" (1996: 21). It

[4] Different characteristics of test situations refer to a test that takes in different time, places or by different delivering modes.
[5] TLU refers to target language use. In different TLU domain, the TLU tasks can be different. Such as the TLU domain of the TOEFL test is the academic settings in colleges and universities in English speaking countries, the TLU tasks should then be within the this domain.

indicates that in order to measure test takers' language ability in similar language use situations, the design features of test tasks should be made as similar as possible in order to ensure the test reliability.

2.1.2 How to make a test reliable

In order to design and develop a reliable test, Hughes (2001, 38) suggests the following ways. He mentions the influential factors to the test reliability, including number of questions, language of items and instructions, ways and reasons to make the test familiar to test takers, the test organization, and the scoring method.

- There should be as sufficient questions in a test as possible. The more the questions or the test tasks are, the more possible that test takers' performance will be similar. Items or questions that are used in a test should be able to distinguish test takers from stronger ones to weaker ones. The more varied the questions are in difficulty levels, the more clearly test takers are distinguished in their scores, and then the more reliable the test would be.
- All the items should be clear in being right or wrong, so that test takers' performance can be distinguished clearly. If some items were designed to be wrong ones, but turn out to be right, the score of the test is doubtable in its consistency to test takers' language abilities and the reliability of itself at the same time.
- Instructions of test task should be provided clearly and explicated. It is very obvious that in order to show their language abilities, test takers should know very clearly what they are asked to do in the test. Test takers should not be misled or confused by the instructions. Without this precondition, the test cannot be reliable in its result. In addition, tests should be well laid out and legible.
- Making test takers familiar with format and testing techniques before they take the test helps test takers to perform well enough and to show their real language ability without influencing by factors not related to their language ability. Distribution of sample tests and practical materials for the test can make a test to be more reliable in this way.

- The administration of the test should be beneficial for test takers, not being disordered or distracting. Noisy surroundings and a disordered organization will affect test takers' performance in a bad way.
- Objective scoring questions, such as Multiple-choice questions, are recommended in terms of reliability, because scoring standards of this kind of questions are clear and perfectly the same to every test taker. The score of a multiple-choice question is more reliable in comparing with a score of an open-ended question, because there are many uncertain factors that will affect the examiners during the scoring process, i.e. their personal preference and social background.
- To ensure the scoring standard is to the largest extent unified, scoring keys for each question should be provided as detailed as possible and scorers should be trained.

All these tips can be used not only to guide the test design, but also to assess the test reliability as well. Since this essay concerns the reading comprehension test reliability, theories and primary research in the following sections will focus on reading from three aspects, which can affect the reading test reliability. They are: how to identify reading ability; how to test reading skills; and how to score reading test.

2.2 Identifying reading ability

Traditionally, reading ability was considered to be one of the four basic abilities in learning a foreign language. The other three are speaking, listening and writing. Before testing reading ability, it should be made clear first that reading ability has different ranges for different purposes. For many non-English learners and test takers, one main reason and purpose for learning English and attending English proficiency tests is to achieve higher education. Reading ability tests for the academic purpose are then designed for measuring test takers' protential abilities to understand and to learn from scientific textbooks.

Among all the English proficiency tests for academic purposes, TOEFL is a typical one. According to the official TOEFL website, "academic purpose" and "academic setting" refer to related activities in college or university classrooms. The academic purpose of the TOEFL test is clearly presented in one of officially issued TOEFL documents, *TOEFL IBT Performance Feedback for Test Takers*. Accoring to this document,

reading skills in the TOEFL test is identified and rated by standards, such as "understanding academic texts in English", and having "a good command of academic vocabulary and grammar structure". The word "academic" indicates that "vocabulary" and "grammar" tested in the TOEFL reading tests are used or tend to be used in learning and teaching activities in college classrooms.

Based on the primary research published in the *TOEFL Monograph Series*, there are four purposes of reading in academic settings. In the following part, there will be an introduction of each of the academic reading purpose, including the definition and skills involved in dealing with each of them. Finally, questions from the TOEFL IBT and PBT sample reading sections are used to exemplify how these reading purposes can be designed into questions to test the reading comprehension ability.

2.2.1 Reading to find information

"Reading to find information" is one of the basic purposes of reading in academic settings, aiming at locating and understanding discrete pieces of information. Skills involved in "reading to find information" are rapid identification of words, short period memory, and fast reading speed. In order to test this ability, test takers are often asked to search for or match certain separated information. Questions that require basic comprehension to single words and expressions are also of this type.

An example of "reading to find information" in the TOEFL IBT sample test of reading is question 7, asking test takers to pick out the false statement among three other true ones. The true statements can be found in line 29-31 to match item a, b and d. Then item c is quickly picked out as the answer. In this kind of questions, test takers are not asked to understand exactly the meaning of the items and the related paragraphs. This is the difference between reading to find information and reading for basic comprehension, which is when test takers should understand the basic meaning of sentences and paragraphs.

2.2.2 Reading for basic comprehension

Basic comprehension is also a basic purpose for reading as well as a basic test task in reading test. In testing this ability, test takers are always asked to select from given items the main idea or main point of a paragraph or a whole passage. Test takers should be able to read fast and efficiently, focus on useful information immediately and then interpret the meaning of the useful information. In *TOEFL 2000 Reading Framework*, there is a clear distinction between "reading for basic comprehension" and reading to search information:

> Reading for basic comprehension involves understanding a subset of individual ideas, primarily those tied to the thematic content. … Comprehension of information that is not central to the main ides is unlikely to be required for this reading purpose (*TOEFL 2000 Reading Framework*, 1999: 5).

An example of reading for basic comprehension from the TOEFL IBT sample test is question 4, which asks test takers to pick out the way in which scientists determine that a large meteorite had impacted Earth. In answering this question, test takers look back to paragraph 3. In line2-3 of this paragraph, there is a sentence: "Scientists first identified this impact in 1980 from the worldwide layer of sediment deposited from the dust cloud that enveloped the planet after the impact." If test takers can understand this sentence, they may choose the answer correctly, which is B: "They found a unique layer of sediment worldwide." The difference between reading for basic comprehension and reading to find information is that multiple-choice items for reading to find information are always directly quoted from the text; items for reading for basic comprehension are always re-written.

The first two purposes of reading in the academic setting have been used in TOEFL IBT and PBT for many years. However, the next two purposes are more and more paid attention to and put into practical question settings in recent years.

2.2.3 Reading to learn

The purpose of "reading to learn" in these tests is to learn from the information conveyed by texts. "Reading to learn" tests involve test takers' abilities to develop basic comprehension of ideas. Test takers are required to identify basic logical relationships, such as cause-and-effect relationship, comparison and contrasts relationship, and classification relationship among different information in academic texts. Reading to learn requires the re-organization of information, the ability to organize "conceptual information and to understand the author's rhetorical intent" (E. Mary, William, Keiko, Peter, Patricia, S. Mary, 2000: 6). This quotation indicates that readers and test takers are asked to make various kinds of information into pieces and to draw links between them by their relationship. Tasks that require "reading to learn" may involve interpreting rhetorical intent, making inferences, and distinguishing relationship between pieces of information conveyed by rhetorical expressions.

In the IBT TOEFL reading sample test, question 2 and 3 are examples of this kind: question 2 asks test takers to identify the reason why the author include certain information in paragraph 2; in question 3, test takers are asked to identify the inferred information from a piece of given information. These examples indicate that answering "reading to learn questions" involves basic comprehension ability and logical re-organization ability at the same time.

2.2.4 Reading to integrate information across multiple sources

To intergrate information across multiple texts is a comprehensive ability, involving all the other three purposes above. It requires readers and test takers to work on no less than two texts from differnce sources, interweaving pieces of information, organizing and making use of complexed information.

Question 13 in the TOEFL IBT sample test of reading can be categorized as this kind. In question13, a sentence is pre-provided, and test takers are asked to interweave it into a paragraph. The pre-provided sentence is a new source or a new piece of information, so this question is testing the ability to deal with integrate of information across different sources. Test takers should locate the concerned sources, understand the basic meaning of the sources and identify the relationship among the different sources before intergrating the different sources. Thus, this kind of task requires more comprehensive

ability, involving the other three abilities above, i.e. reading to find information, reading for basic comprehension, and reading to learn.

2.3 Testing reading skills

As aforementioned, with different purposes in different situations, the range of reading ability can be varied. Reading-skill testing should thus be based on the identification of reading tasks and reading ability. Generally speaking, there are three factors that can affect the reading test process and result: test setting, test difficulty, and scoring method.

2.3.1 Test setting

Test setting consists of several external features of a test, including physical setting, test delivering mode, time allotment, and the instruction. From these features, what a test looks like is described. Additionally, as external factors, these following features also influence more or less the difficulty of a test as well as test takers performance in the test.

- **Physical setting**

According to Bachman and Palmer (1996, 48), the characteristics of **physical setting** [6] describe the environment in which a test is taking place, including where the test site is located and whether the materials and equipments used in the test, such as pencils, paper, audio-visual equipment, are familiar to test takers.

The reason for considering physical setting is that Bachman and Palmer believe that physical settings can "clearly affect an individual's use of language"(1996: 48). When test site and time can be flexibly and test delivering mode is familiar to test takers, they would feel comfortable, thus possibly perform better.

- **Test delivering mode**

[6] According to Bachman and Palmer (1996), physical setting includes not only those mentioned in the essay, but also other things, i.e. the temperature, humidity, light conditions of the test sites. Since these conditions are generally similar in the IBT and the PBT of the TOEFL test, they are omitted in this essay.

Test delivering mode, which is diverged nowadays in language proficiency tests, refers to the medium by which test is presented in front of test takers. Basically there are two test modes that are co-existing: computer-based (CB) test and paper-based (PB) version (Alderson, 2000: 96). The latter one is traditional and the computer-based test has been improved in the TOEFL test as the internet-based test so as to delivering the test via internet in visual and aural modes at the same time. For technical reasons, the computer-based or internet-based tests can not replace the paper-based versions. As a result, as Khalifa and Weir also mention in *Examining Reading*, these two modes will co-exist for some time (2009: 281).

Khalifa and Weir enlist some research by other testing researchers on scoring and process differences between CB and PB: Mayes, Sims and Koonce proved in an investigation that the more mental workload one experienced, the poorer his/her performance in a test would be; a study by Noyes, Garland and Robbins proved that there is more mental effort involved in CB than in PB in completing reading comprehension tasks. The two primary studies infer, at least theoretically, that the scores of reading tests can be negatively affected by the CB versions rather than by the PB versions. The effect of test delivering mode in testing process and scoring methods will be discussed in terms of TOEFL IBT and PBT in the following sections. Apart from the test delivering mode, which affects the test difficulty and test takers' performance, time allotment for a test also matters.

- **Time allotment**

Time allotment refers to how long the test takers use in finishing each part and the whole test. There can be speeded tests, in which not all the test takers are expected to manage to finish all the tasks. There are power tests, in which enough time is given and every test taker will be able to complete every task or answer every question. Speeded tests tend to be less difficult than power tests because in speeded test, the language of texts and the difficulty of questions tend to be easier and test takers' reaction should be quick. Additionally, when the time span is longer for the whole test, the test tends to be more difficult because longer-time test is more exhausting and test takers should make more efforts to concentrate in the test.

- **Instructions**

Instructions should be made as explicit and clear as possible in informing the steps for test taking and scoring. Language in instructions can be test takers' native language, or target language, or even both. Instructions can be presented in aural, visual, or both ways. Instructions can also be of very different forms: lengthy or brief; with or without example; provided all in one time or provided dividedly in front of each new question. Brief instructions, in native language and with an example, are considered the easiest to follow. It is disputable whether the aural way or the visual way in presenting the instruction is easier for test takers to follow. Some researchers argue that instructions and test directions in both aural and visual ways at the same time can help test takers to understand the test directions better (Alderson, 2000: 205). However, what matters is not the way instruction is provided, but whether it is provided explicitly and to be easy to understand.

2.3.2 Test difficulty

Test setting is considered as the external factor of a test, which concerns the "form" of a test. In comparison, test difficulty is the internal factor of a test, which concerns the "content" of a test. Test setting only *affects* test reliability, but test difficulty *decides* test reliability. According to Alderson (2000: 206), the difficulty of reading test is mainly affected by three decisive factors: the difficulty of the text/passages, the difficulty of questions and the relationship between questions and the related text. Besides these three main factors, there are other influential factors, such as text length, whether the text is present when test takers answer questions, and whether questions are independent to each other. Their influences on test difficulty are briefly discussed in this section.

- **Text length**

Text length is the most obvious aspect to compare between one text and the other. Generally speaking, the longer the text is the more questions that are followed, thus longer texts take longer time to read and to answer questions. However, when texts are shorter, there can be more passages in the whole test, which cover wider topics and require more background knowledge. Therefore, test difficulty cannot be simply judged by the lengths of texts. Text and question difficulties and their relationships then need to be analyzed.

- **Text difficulty**

In analyzing **text difficulty**, the text material analysis method will be adopted. It is presented in *TOEFL 2000 Framework* (Jamieson, 1999: 18-30) that text materials can be analyzed from three aspects, i.e. grammatical features, discourse features, and pragmatic features. In this section, grammatical features will be used in identifying text difficulty.

Grammatical features refer to syntax of the sentences and vocabulary used in a text. Jamieson and his co-researchers believe that words and phrases used in daily life are easier to understand than academic ones; and specific words and expressions in certain subjects, involving background knowledge while reading, are more difficult to comprehend than words and expressions that can be easily understood.

Among the four sentence types in English, i.e. simple sentence, compound sentence, complex sentence, and compound-complex sentence, complex and compound-complex sentences are usually difficult to understand. Because they are combined and formed by simple sentences and clauses (*Milton's Grammar*, 157), complex and compound-complex sentences usually contain more information than simple sentences do. But good writings, such as those selected to be test texts, should not contain only simple sentences or complex sentences. Well-written texts always use various types of sentences, among which compound-complex sentences are the most difficult to comprehend; and simple sentences are relatively easy to understand.

Grammatical features will be used in the next section in analyzing and comparing the text difficulty between the IBT and the PBT sample reading tests. Besides text length

and grammatical features, the question difficulty and the presence of text while answering questions also affect the reading test difficulty (Alderson, 2000: 202).

- **Question difficulty**

In analyzing **question difficulty** in a test, language use, length, and question type are three aspects that should be brought into investigation.

a. The vocabulary and syntax difficulty of questions: many reading assessment researchers agree that in well-designed reading tests, words and sentences used in questions and items[7] should be simpler than those in the text (Bachman, Alderson). Academic words and complex-structured sentences are likely to lift the difficulty of the question.

b. The length of questions and items: In comparing with long questions and items, short questions and items take less time to read. If the short questions and items in test A are more than in test B, A test takers are likely to have more time to think about how to answer the questions. As a result, they possibly get higher scores.

c. Multiple-choice question: Although some researchers, such as Jack Upshur, have questioned the phenomenon, multiple-choice question is the commonest way of assessing reading. According to Munby (1968: 25) and Alderson (2000: 204), the reason is that multiple-choice questions can show what the test takers' misinterpretations are for some parts of the text. Therefore, the result of multiple-choice questions can be used as statistics for reading assessment research and to guide reading test design for next time.

The difficulty of questions is affected by the language use and length of the question and items. But the relationship between questions, items and the text is the most crucial factor that decides the difficulty level of questions.

- **Relationship between question and text**

[7] Items here refer to answers below multiple-choice questions that are provided for test takers to choose.

Besides “text difficulty” and “question difficulty”, **the relationship between questions and text** is considered by Alderson as the crucial factor that decides reading test difficulty (2000: 113). When the four reading purposes in academic setting, mentioned in section 2.2, are used in designing questions for a text, questions for testing different reading purposes are of different difficulty levels. Hereafter, the four types of question and text relationship will be interwoven with four reading purposes in order to identify how the relationship between question and text affects the reading test difficulty.

a. Whether the questions and answers are based directly on texts. Alderson claims that questions that are explicitly based on the text tend to be easier than those are not. That is to say, if the answer for a question can be found directly from the text, the question is considered to be easier than the answer for which is not the original sentence from the text. For example, questions for reading to find information and questions for basic comprehension are explicitly based on text, but questions for reading to learn and for reading to integrate information are not directly based on information in the text. That is why the latter two kinds of questions are more difficult than the former two kinds.

b. Whether the location of the answers in the text is in the same or varied places. As we can say from our own test-taking experience, if the location of the answer is simple, in a sentence or in one paragraph, it takes shorter time for us to find it. For this reason, questions for reading to integrate information across multiple sources are more difficult than questions for the other three purposes, because in answering questions for information integration, test takers at least have to read two sources of information, which are usually in different places.

c. Whether background knowledge is involved in question answering. Questions, in answering which requires test takers background knowledge, are more difficult than those are only based on information in the text. The reason why questions for reading to find information are more difficult than questions for reading for basic comprehension and those for reading to learn is that the former one needs no background knowledge, and test takers simply find information from the text.

d. Whether the questions are designed according to the four purposes of reading in academic settings. In the same passage, questions for testing the first purpose are the

simplest ones, while questions for testing the last purpose are of the highest difficulty level (see Section 2.2). Table 3 in the following shows the difficulty relationship among questions that are designed for the four purposes of reading.

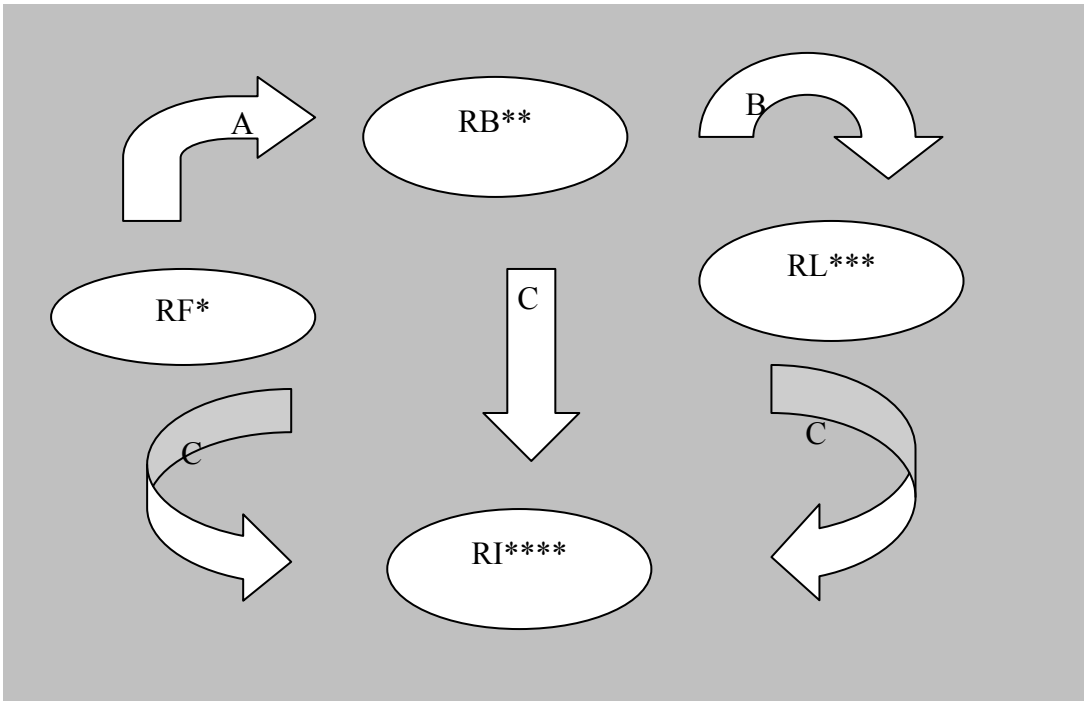

Table 3 Reading purposes and question difficulty
*RF refers to "reading to find information"
**RB refers to "reading for basic comprehension"
***RL refers to "reading to learn"
****RI refers to "reading to integrate information".

Questions that are designed for the four purposes of reading are considered as of different difficulty levels: questions for the purpose in the circle, which is pointed at by the arrow, is more difficult than questions for the purpose at the end of the arrow, and to answer the question at the end of the arrow involves the ability in answering questions at the beginning circle of the arrow. For instance, Arrow A shows that questions for "Reading for basic comprehension" are more difficult than questions for "Reading to find information", and in answering questions for basic comprehension ability of reading, test takers should be able to find the related information first. There are three arrows named by C, all pointing at "Reading to integrate information". They show that questions that are testing "Reading to integrate information" are the most difficult ones because to answer them requires finding relevant information, understanding the basic meaning of the information and to identifying the logical relationship among the different pieces of information.

As demonstrated before, the relationship between question and text is the crucial factor for test difficulty identification. a, b, c, and d above can be comprehensively interpreted as the following so as to get a conclusion to text-question-relationship and reading test difficulty: the difficulty level of questions is: PF<PB<PL<PI: questions designed for testing PI are the most difficulty; questions for testing PF are relatively easiest one (see p. 37).

- **The presence of text while answering questions**

Alderson (2000) claims that, besides the length of texts, **the presence of text** when questions are answered and the independence from one question to another will also affect the difficulty of a reading test. According to Alderson, if test takers are allowed to read the text again when they are answering questions, they are likely to get higher score. He also points out that in paper-based tests, test takers are often allowed to review texts when answering questions. But in computer-based test, things may be different. He seems to state that if a computer-based test provides no opportunity of reviewing, it can be more difficult, in this sense, than a paper-based test. Take TOEFL for example, if in the IBT test takers are not allowed to review the text, the difficulty of its reading test can be considered higher than that in the PBT, because in the IBT, test takers have to use short memory and note-taking during reading test, while in the PBT, they do not have to.

2.4 Scoring reading tasks

In order to measure test takers' reading abilities, the concept of the scoring method needs to be introduced. Since the score is the main goal for test takers and scores are concrete in numbers and thus comparable, the reading test scores are always considered to present test takers' reading comprehension abilities. For this reason, the score reliability is the crucial factor to ensure the test reliability.

2.4.1 Scoring method

Scoring method directly relates to the purpose of language test task, which is to measure the test takers' language ability. The evaluation of the test takers' performance is carried out by a scoring method, "which specifies how numbers will be assigned to test takers' performance" (Bachman and Palmer, 51).

Three methods for scoring are: criteria for correctness, which determines how the correctness of the response or answers by means of an objective scoring key, multiple value rating scales, or judgments of correctness/incorrectness; procedures for scoring the response, which refer to the steps that make out the scores; and explicitness of criteria and procedures, which refers to whether or not or to what extent information on how the test will be scored is given, omitted, or made vague deliberately. Bachman and Palmer also demonstrate that objective questions, which involve no examiners' subjective influence, ensure the score reliability.

2.4.2 Score reliability

As mentioned on a free research website[8], "the reliability of a score is an indication of how much an observed score can be expected to be the same if observed again." When the same test, or the similar test paper are given again, and if the score of same group of test takers show that they have the same level of language ability, the score is then considered reliable. "Same level" should be highlighted because within a test, there can be more than one delivering modes, and the scoring systems can be different between different versions. Thus, the comparison between the "numbers" of scores accounts for nothing. In the case that more than one scoring systems existed in one test, the reliability of the test relies on whether scores, which are achieved in different scoring systems, indicate that the same test takers are of the same "level" of language ability. If so, the test results can be considered to be reliable.

3. Analysis and discussion

[8] To learn more about score reliability, access http://www.chinazzwb.com/news-abroad/12241495.html [Accessed 12th April, 2010].

TOEFL IBT and PBT are of obviously different modes: the PBT is delivered in paper and the IBT is delivered by the internet and computer. Since the IBT was introduced in 2004, ETS has been publicizing and advocating the IBT all over the world, and ETS wants the IBT to be the major mode of the TOEFL test (*Bulletin for PBT of TOEFL in 2010-2011*, 2009: 3). However, in some countries and areas, for technique reasons, the PBT is still the only way for people to take the TOEFL test.

As mentioned by the TOEFL bulletin of the IBT and some testing researchers (Alderson, Khalifa, and Weir), the co-existence of the IBT and the PBT will definitely continue for some time in the future. As the pre-condition for the analysis and discussion in this section, the co-existing status of IB and PB versions [9] in the TOEFL test makes it necessary and useful to analyze the consistency of two versions in the name of test reliability. And as aforementioned, demonstration of the TOEFL organization (see p. 2) claims that same results can be yielded from the IBT and the PBT. Then in this section, three influential factors for test reliability will be analyzed in order to find out whether the TOEFL test is reliable within the reading comprehension tests in the two different versions.

3.1 Reading test setting in TOEFL

According to theories concerning test settings, the analysis of TOEFL IBT and PBT test settings will be carried out by physical setting, delivering mode, instruction, structure, and time allotment. The analysis in this section is based on documents issued on the official TOEFL website, especially in *Bulletin for IBT of TOEFL in 2010-2011* and *Bulletin for PBT of TOEFL in 2010-2011*. Hereafter, these sources will not be repetitively mentioned.

3.1.1 Test setting in the IBT

[9] IB refers to internet-based; PB refers to paper-based. They are two co-existing delivering version of test nowadays.

The reading comprehension test in the IBT measures the ability to understand academic reading material. Instructions in the IBT reading test are visual on screen, including two parts: instructions for reading test and instructions for computer manipulations. There are 12 lines in "Reading section Direction". In reading and answering questions in each passage, 20 minutes are given. The reading and answering process is timed by computer system automatically. When time is up, test takers will not be allowed to stay at the same passage or to answer questions. The next passage will show up. The IBT is of computer-based delivering mode. All the passages and questions and instructions are delivered on computer screen. Test takers use computer and mouse in reading texts and answering questions during the test. The test place, which is called "the testing site" in the IBT TOEFL Bulletin, is pre-fixed by the test organizer.

The following chart is an excerpt one from the IBT test description, which is also published by the official TOEFL website:

Section	**Description**	**Testing Time**	**Questions**	**Score Scale**
Reading in **IBT**	3-5 passages from academic texts; approximately 700 words long; 12-14 questions per passage.	60-100 minutes	36-70 questions	0-30

Table 4. IBT reading test description*

*This chart is a part of test setting description issued on the official TOEFL website. In the original table, there are four "Sections", including reading, speaking, listening and writing.

Table 4 indicates that in the IBT reading test, test takers have 20 minutes in reading one text and answering all the questions for it. The average speed limit is about 35 words per minute and 84-90 seconds for each question. In other words, every minute, averagely, test takers should go through 3 lines, answering 2/3 of a question.

3.1.2 Test setting in the PBT

The reading Comprehension test in the PBT measures the ability to understand non-technical reading material. The PBT is offered six times a year in specific locations, which refer to countries and areas without internet connection or possibility to use computer. TOEFL PBT is approximately three and one-half hours long, in which

reading comprehension test takes 55 minutes. During the whole time span for the reading comprehension test (55 minutes), test takers are asked to finish answering all the 50 questions. The PBT is of paper-based delivering mode. Test takers use paper and pencil in reading and answering questions during the test. The PBT tests are taken place in certain pre-fixed places chosen by test administrations.

The following chart is an excerpt one from the IBT test description, which is also published by the official TOEFL website:

Section	**Description***	**Testing Time**	**Questions***	**Score Scale**
Reading in **PBT**	5 passages from non-technical texts; approximately 350 words long; 10 questions per passage.	55 minutes	50 questions	31-67

Table 5. The PBT reading test description

*when originally published in the official TOEFL website, there are no "Description" and "Questions" in the table. In making the PBT test description comparable to the IBT, the writer of this essay completed the two items by consulting the PBT former test papers.

Table 5 indicates that in the PBT reading test, test takers have 11 minutes in reading one text and answering all the questions for it. The average speed limit is about 32 words per minute and 66 seconds for each question. In other words, in every minute, averagely, test takers should go through almost 3 lines and answer one question.

3.1.3 Discussion of the TOEFL test setting

Time and place for the IBT are more flexible than that for the PBT: test takers can choose one time among 60 provided ones and different test sites are provided by the TOEFL test administration. The flexibility in time and place benefits the IBT test takers by providing them a nearest place and a proper time to take the test. The physical setting of the IBT is more humane than the PBT.

In the IBT, reading and answering process for each passage is timed by computer. As a result, the IBT test takers cannot allocate time to each passage by themselves. The good thing for the strict speed limitation in the IBT is that it ensures the fairness. Each test

takers use the same time in dealing with each passage. Those who fail to do well in one passage still have the equal chance to perform in other passages.

The PBT, on the other hand, allows test takers to allocate time by themselves. Within the whole time limit in reading comprehension test, test takers can go forward to read next passage as they like. They can also use much more time on a certain passage when they think it necessary. In the PBT, if a test taker finds one passage is much more difficult for him, he can use more time on it in order to answer questions for the passage better. The good thing for the relatively loose time allotment is that in order to perform better, individual test takers can use more time on passages or questions they find more difficult.

Test descriptions are different. Table 4 and 5 show that there are two more items in the IBT description than in the PBT (see* in Table 5), including the exact number and length of passages and the number of questions after each passage. Thus, test takers should consult sources to find such detailed information about the PBT reading test.

Time allotments for the IBT and the PBT reading test are different, showing comparatively in the following chart:

	Time	**Passage numbers**	**Passage Length**	**Question numbers**	**Speed limit/ per passage**	**Speed limit/ per minute**	**Speed limit/ per question** [10]
IBT	60-100 minutes	3--5	700 words	12-14	20 minutes	35 words	84-90 seconds
PBT	55 minutes	5	350 words	10	11 minutes	33 words	66 seconds

Table 6. A comparison between reading speed required in the IBT and the PBT test

[10] Even though questions in the IBT reading provide a bit more time to answer, it is not sufficed to say that the IBT reading is easier. Actually, a bit more time provision indicates that questions in the IBT reading test are more difficult to answer. Analysis on the IBT and the PBT questions is in Section 3.2 that follows.

Table 6 shows that, to compare with the PBT, the time span of the IBT is longer, the length of each passage in the IBT is 100% longer and the questions after each passage in the IBT are 20%-40% more than that in the PBT. Even though the time for each passage is longer in the IBT, the reading speed it requires is faster.

Research by Noyes, Garland and Robbins (Alderson, 2009: 281) revealed that computer or internet, as testing delivering modes, have a negative effect on reading test results to compare with the paper-based version. However, the study above is only theoretically true. In the specific case of the TOEFL test, the statistics collected from the IBT and the PBT show that in reading test setting aspect, these two versions are very different, not only in delivering modes, but also in all the other factors only with the exception of "structure".

Physical Setting	IBT: Test time and test sites can be chosen by test takers for their convenience. A lot of time and places are enlisted for chosen.
	PBT: Only six times per year; Test sites are very fixed.
Mode	IBT: Passages and questions are delivered by computer and internet; Test takers use computer and mouse to read passages and answer questions
	PBT: Passages and questions are delivered on paper. Test takers use paper and pencil to read and answer.
Instruction	IBT: Explicitly presented to test takers in five items
	PBT: Rougher and vaguer than IBT, including only three items
Structure	IBT: Latter appeared passages are more difficult; question for the first paragraph is in front of that for the second paragraph.
	PBT: the same to IBT
Time Allotment	IBT: computer-controlled and each passage is of 20 minutes
	PBT: the whole time is controlled but test takers can adjust time for each passage by themselves.

Table 7. A comparison between the IBT and the PBT test settings

After the analysis of test settings of the IBT and the PBT, it is not sufficient to say either of the test settings is crucial enough for the test to be reliable and the five aspects

in table 7 are not decisive factors for the IBT and the PBT to be different. They are only influential factors that can make test process easier or harder for test takers. With the exception of the "structures" in which, both the IBT and the PBT are designed similarly, both the IBT and the PBT settings have positive factors and negative factors in affecting their test results: the convenient physical setting and the explicit instruction of the IBT make it easier for test takers; time allotment makes the PBT easier than the IBT because time span in the PBT is as a whole shorter than in the IBT and time can be controlled by test takers in the PBT also positively affect the PBT test result. Delivering mode is most obvious difference between the IBT and the PBT, and the influence of delivering mode depends on test takers' computer and internet manipulating abilities: in developed regions and countries, test takers tend to be quite familiar with computer and internet, the IBT then can be convenient and made easy to attend. However, test takers' performances can be negatively influenced in regions and countries where computers are not widely used or internet access is not widely available. So test setting is not the very crucial factor for testing reading ability. We now move on to the analysis of reading test difficulty.

3.2 Reading test difficulty in TOEFL

In this section, there is an analysis on text and question difficulties in the IBT and the PBT reading test. The analysis will be carried out by using sample test papers (see Appendix A, B) for the two versions respectively. According to relevant theories for test difficulty analysis, each subsection includes three parts: text difficulty, question difficulty, and relationship between text and questions. Finally in this section, there will be a comparison between the IBT and the PBT in their test difficulty.

3.2.1 Test difficulty in the IBT

According to the relevant theory on test difficulty in section 2.3.2, in this section and section 3.2.2, the IBT and the PBT test difficulties are analyzed from the following five aspects: the presence of the text while answering questions; text length; text difficulty; question difficulty, and relationship between question and text.

- **The presence of the text while answering questions**

In the IBT, "when the time is up, the passage will erase and the first question will be shown" (*Bulletin for IBT of TOEFL in 2010-2011,* 2009: 13) It means that if test takers answer questions within the time limit, the text is present and available to be consulted. When reading time is up, only questions are presented, with no text any more. The strict time limit can make test takers nervous and feel pressure. In this sense, the IBT reading test is made more difficult.

- **Text length**

The reading text in sample test for the IBT is 57 lines long, with 7 paragraphs and 14 questions afterwards. There are approximately 960 words in the passage, which is a bit longer than the general length of about 700 words, which is regulated in the test descriptions (see Section 3.1.1).

- **Text difficulty**

According to *TOEFL 2000 Framework* (Jamieson, 1999), analysis of the IBT reading sample text difficulty will follow two aspects:

***Vocabulary features*.** The title of this passage is "Meteorite Impact and Dinosaur Extinction", implies that words used in this passage are possibly related to the biological evolution of the Earth. Vocabulary in the biological evolution group is very academic, because people seldom use words in this group, such as "Cretaceous" and "paleontologist" in their daily life, but most possibly in classroom during teaching, learning or testing process. There are some academic subjects mentioned in the passage, i.e. "biological", "ecological", and "geological". In the following chart, there are statistic figures showing the number of words and phrases that are related to the three aforementioned scientific subjects:

	Biology	Ecology	Geology	Total
Words	9	8	12	29
Phrases	6	11	10	27
Total	15	19	22	56

Table 8. Numbers of academic words and phrases in the IBT sample reading test

Numbers of scientific words counted in table 8 are not overlapped with each other. The table shows the number of newly appearing academic words and phrases in the IBT sample reading test. It is not sufficient to say how difficult the text is academically. Calculating from the total number of academic words and phrases, every line of the text contains one new academic word or phrase. As known from the former section analyzing results that in the IBT reading, test takers have to go through about 3 lines per minute. Therefore, we now know that, averagely speaking, test takers encounter 3 new academic words or phrases per minute.

Actually, the amount of academic words and phrases that test takers will encounter is much more than 3 per minute, because in the text, the words and phrases appear repetitively. Table 9 in the following shows specifically how many academic words and phrases test takers will encounter during their reading process.

	Academic Words	Academic Phrases	Total Number of times
Paragraph 1	3	4	7
Paragraph 2	10	7	17
Paragraph 3	10	13	23
Paragraph 4	11	14	25
Paragraph 5	4	5	9
Paragraph 6	10	5	15
Paragraph 7	4	1	5
Total number of times	52	49	101

Table 9. Times of academic words and phrases appearing in the IBT sample reading text

Table 9 shows the frequency of academic words and phrases appearing in the IBT sample reading test. The total number of times for each paragraph indicates that the first and last paragraph contain the least academic words and phrases; paragraph 3 and 4 contain the most academic expressions. Thus, from the vocabulary aspect, paragraph 3 and 4 in the middle of the text should be the most academically difficult paragraphs. How about their sentence difficulties? Do those paragraphs contain more academic expressions or have complex sentence structures? Answers should be based on the analysis on syntax features of the text.

Syntax features. Theory on sentence types and their difficulty levels indicate that paragraphs containing more complex sentences and compound-complex sentences tend to be more difficult to understand. The following chart shows the number of sentences in each paragraph in the IBT sample reading text, and sentence types within each paragraph:

	Number	Simple	Compound	Complex	Compound-complex
P. 1*	3	2	0	1	0
P. 2	6	3	0	1	2
P. 3	6	2	0	3	1
P. 4	5	2	0	2	1
P. 5	4	0	2	2	0
P. 6	5	4	0	1	0
P. 7	4	1	1	1	1
Total	33	14	3	11	5

Table 10. Figures about sentence types in the IBT sample reading test

*P. refers to Paragraph in the sample reading test text

The figures in table 10 indicate that paragraph 2, 3, and 4 are possibly of the highest difficulty level for test takers to understand: these three paragraphs contain more complex and compound-complex sentences than the other four paragraphs do. Combining with the analysis on vocabulary in the text, paragraph 3 and 4 are of the highest grammatical difficulty level.

As mentioned in section 2, the test difficulty is not only decided by text length and language use, but also crucially affected by the difficulty of questions and items. Hereafter there will be analysis on question difficulty of the IBT sample reading test.

- **Question difficulty**

Alderson (2000: 113) provides the following two aspects in analyzing question difficulty in a reading test, including the vocabulary and sentence use in questions and item lengths in questions.

***Vocabulary and sentence use*.** Language of questions in the IBT sample reading test is clear and explicit—all of them are less than two lines long. Words in questions are no more difficult than words in the text. There are no new academic words that appear in the text, in questions. All academic words in questions are all quoted from the text, which indicates the language of questions does not lift the test difficulty.

In the 14 questions, question 1, 5, 6, 8, 9 have no more than twelve words, within one line. All of the five questions, asking for the meaning of a specific word in the text, are quite easy to follow. Question 5 is an example:

5. The word "excavating" on line 25 is closest in meaning to
a. digging out
b. extending
c. destroying
d. covering

Table 11. One question excerpted from the IBT sample reading test

It is obvious that in question 5, all the other words are simpler than the one which is tested, "excavating". This fact proves that the language of questions does not make the question more difficult to answer.

Except the last question, the rest eight questions 2, 3, 4, 7, 10, 11, 12, 13 are as long as about two lines. Half of them are simple questions, with no clause in them. The other four have clauses in them, which will be taken longer time for test takers to read and to

understand, thus they lift the difficulty of the whole test. The last question is of six lines long. The language of question 14 is much simpler than language in the text. But the length of it still takes time and lifts the difficulty of the whole test. Therefore, the language of five questions lifts the difficulty of the whole test.

Item length. Items in question 1, 5, 6, 9, 12, and 13 are of one single word. It implies that test takers can read these items by a glance while they are reading the relevant questions. Thus, these items will not lift the test difficulty. Question 8 also has short items, which are only of about three words long. So there are in all seven questions with simple enough items, which do not lift the test difficulty. Items in the other seven questions are complete sentences about two lines long. To read and to understand those items cost test takers time and misunderstanding of the items will directly result in lost points. Therefore, items in the other seven questions lift the test difficult.

- **Relationship between question and text**

According to the influential factors concerning relationship between question and text in section 2.3.2, as Alderson's states that the test difficulty is crucially affected by the relationship between questions and the text (*TOEFL 2000 Reading Framework: A Working Paper,* 118). Question and text relationship in the IBT sample reading test will be presented in the following chart. The difficulty of each of the 14 questions will be identified by their relationship with the text and the reading purposes (see Section 2.2, and Table 3 in Section 2.3) that they are designed for.

	Q*1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q 9	Q 10	Q 11	Q 12	Q 13	Q 14
PF**	√***	≈	≈	≠	√	√	√	√	√	≠	≠	≈	≠	≠
PB	≠	≠	≠	≠	≠	≠	≈	≠	≠	≈	≠	√	≈	√
PL	≠	√	√	√	≠	≠	≠	≠	≠	√	√	≠	≈	≠
PI	≠	≠	≠	≠	≠	≠	≠	≠	≠	≠	≈	≠	√	√

Table 12. The relationship between questions and reading purposes in the IBT

*Q: question;
**P: reading purpose: PF: reading to find information; PB: reading for basic comprehension; PL: reading to learn; PI: reading to integrate information.

***√: Q is directly designed to test P; ≈: Q is partially designed to test P; ≠: Q is not designed to test P

Besides the most crucial decisive factor in test difficulty, namely the relationship between questions and text, other factors, as analyzed before in this section also affect the IBT test difficulty to some extent. The same happens in the PBT test. In the next section, the PBT reading test difficulty in the same pattern is analyzed as in this section. After the PBT test difficulty analysis, there follows a comparison between the reading test difficulties between the internet-based and paper-based versions of the TOEFL test. To what extent the two versions are consistent in test difficulties is essential for the reliability of the TOEFL test.

3.2.2 Test difficulty in the PBT

Following the same order and from the similar five aspects as in section 3.2.1, test difficulty in the PBT is analyzed in this section as follows:

- **The presence of text while answering questions**

When test takers are answering questions that follow each passage, they are allowed to re-read the passage and find information. This is very different from the situation in the IBT, as mentioned in section 3.2.1. The presence of text while answering questions can definitely reduce the difficulty of a reading test because on one hand, test takers will be less nervous with the text to consult and to read again, as compared to the situation in the IBT; on the other hand, short memory is not involved while test takers are answering the questions, so they will be more concentrated on the understanding of the text and questions.

- **Text length**

The sample reading test for the PBT is 34 lines long. It is about 410 words in 4 paragraphs, and afterwards there are 10 questions. The length of this sample test is approximately the same as passages in real test papers.

- **Text difficulty**

Vocabulary features. This sample reading test in the PBT, with no title, is mainly about Alaska pipeline. Vocabulary use in this passage is not related to a certain scientific subject. This passage is a non-technical article, containing information about the Alaska pipeline, including its beginning and destination, as well as the construction of it. The following chart shows the numbers of specific words and phrases related to the topic of the text.

	Words	**Phrases**
Paragraph 1	2	1
Paragraph 2	3	0
Paragraph 3	4	1
Paragraph 4	1	1
Total	10	3

Table 13. Vocabulary features in the PBT sample reading text

The specific words and phrases are not used in scientific subjects, thus they are not academic ones. They are selected out only because that they are related to the main idea of this passage. Additionally, many selected words and phrases are quite commonly used in our daily life, such as “canyons”, and “Arctic Ocean”.

Syntax features. Based on the description of English sentence structure, presented in section 2.3.2, the following chart shows the number of sentences in each paragraph and their structures.

	Number	**Simple**	**Compound**	**Complex**	**Compound-complex**
P. 1*	3	2	0	1	0
P. 2	4	1	1	2	0
P. 3	4	4	0	0	0
P. 4	4	3	1	0	0
Total	15	10	2	3	0

Table 14. Syntax feature in the PBT sample reading test *P: paragraph

According to the primary research on the relationship between sentence structure and text difficulty in Section 2, in table 15, 2/3 of all sentences are simple sentences, which can then be considered less difficult than complex and compound-complex sentences in the same passage. When syntax difficulties are compared between two passages, the one with a high percentage of complex and compound-complex sentences is considered to be more difficult to understand.

- **Question difficulty**

Since question difficulty in the IBT has been analyzed in section 3.2.1, here in this part, language use and item length of questions in the PBT are to be analyzed. Comparison between the two will be presented under discussion in the next section.

Vocabulary and sentence use. Questions in the PBT sample reading test are clearly delivered. Questions are short and explicitly delivered: all of the ten questions are less than two lines, and seven of them are one-line long. There are no words or phrases that are newly appearing in questions. So the length and word-chosen of the questions do not lift the difficulty of the test.

Item length. Items following question 1, 2, 3, 6, 7, and 8 are one-word long. As mentioned in "Item length" of the IBT in section 3.2.1, these one-word-long items cost almost no time for test takers to read. In other words, they save time for test takers to read them, so they do not lift the test difficulty. Items for question 4, 5, and 10 are about 2 or 3 words long, which are also time-saving and do not lift the test difficulty. Among all the items for the total ten questions, only items for question 9 are of one-sentence long. To read them takes extra time from the whole time span of the test. So we may say that there is 10 percent of the items here possibly lift the test difficulty, for they are long enough to take extra time to read and to understand and the misunderstanding of the items can lead test takers to lose points in the test.

- **Relationship between questions and text**

The relationship between questions and text in the IBT is analyzed by the relationship between question design and reading purposes. Table 15 in the following, in the same pattern as in table 12, shows which reading purpose questions the PBT sample reading test are designed for.

	Q1*	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
PF**	≠	√	√	√	√	√	√	√	≠	√
PB	√***	≠	≠	≠	≠	≠	≠	≠	≠	≠
PL	≠	≠	≠	≠	≠	≠	≠	≠	√	≠
PI	≠	≠	≠	≠	≠	≠	≠	≠	≠	≠

Table 15. The relationship between questions and reading purposes in the PBT

*Q: question;

**P: reading purpose: PF: reading to find information; PB: reading for basic comprehension; PL: reading to learn; PI: reading to integrate information.

***√: Q is directly designed to test P; ≠: Q is not designed to test P)

Till now, by using the sample reading test papers for the IBT and the PBT, the influential factors of reading test difficulties have all been analyzed. In considering all these aforementioned influential factors, there will be a comparative analysis on the reading test difficulties between the IBT and the PBT.

3.2.3 Discussion of the TOEFL test difficulty

Based on the respective analysis on test difficulties of the IBT and the PBT in terms of reading, there is a comparison in this section between each item that affects reading test difficulties. After the comparison, we can conclude whether the reading tests in the IBT and in the PBT are of the same difficulty level or which one is more difficult.

- **The presence of text during answering process**

In the IBT, test takers' answering process is timed by a pre-fixed computer project. Within 20 minutes for each passage, test takers are allowed to re-read the text while they are answering questions. When time is up, they are still allowed to answer questions but text will not be available to consult. In the PBT, test takers can allocate

their time by themselves, and during the whole testing process, the texts are available to re-read. Therefore, in the PBT, test takers, having more chance to think over and revise their answers with the presence of the text, are likely to get higher scores by. As a result, the PBT tends to be easier in reading test than the IBT.

- **Text length**

There are 7 paragraphs in the IBT sample test and 4 paragraphs in the PBT. Paragraphs are longer in the IBT and there are four more questions in the IBT than in the PBT. Reading speed in the IBT is required to be faster (see Table 7). So the reading test in the IBT is more difficult than that in the PBT in terms of text length, question numbers and required reading speed.

- **Text difficulty**

In the IBT reading text, there are totally 56 academic words and phrases that appear in one passage, and the frequency of academic words and phrases is 10.53% [11]. In the PBT, on the other hand, there are few academic words and phrases. In stead, there are some specific words that are related to the main idea of the passage. These words are much easier and much fewer than the academic words in the IBT (see Table 14).

Sentence structures are varied in the IBT. Among all the sentences in the passage, about 50% are complex and compound-complex sentences. According to *Milton's Grammar* mentioned in section 2, the more complex and compound-complex sentences exist in a passage, the more difficult the passage is. Sentence structures in the PBT passage are mainly simple ones. There are no complex-compound sentences in the whole passage and the frequency of complex sentence is 20%, which is much less than the number in the IBT passage.

The result of the comparison between the IBT and the PBT text difficulty is that, obviously, the IBT sample reading text is more difficult than the text in the PBT.

[11] To calculate the academic words and phrases frequency, the number of words for the whole passage is 960, which is counted in former section; the number of academic words and phrases is 101. According to Table 8, the repetitive academic words and phrases should be counted because to deal with them takes time and then affects the text difficulty.

- **Question difficulty**

All the questions in the PBT are multiple-choice questions. In the IBT, although all the questions are objective questions, question 14 has three correct items. This kind of question tests test takers' comprehensive understanding to the whole passage and their ability to integrate various kinds of information logically. Additionally, former research by Alderson on question and item lengths (see p. 16: b) indicates that questions in the IBT are more difficult: 50% items in the IBT questions are no more than three-word-long; in the PBT, this figure is 90%. In the IBT, 64% of the questions are more than two-lines lone; but in the PBT, only 10% of all the questions are longer than two-lines. Therefore, in terms of question length, the IBT reading test is more difficult.

- **Relationship between questions and text**

According to Enright and her colleagues' research, mentioned in section 2.3, questions for testing different reading purposes are of different difficulty levels.

The following table comprehensively assesses the test difficulty of the IBT and the PBT sample reading test. Numbers are used to stand for the extent of difficulty in order to make the test difficulty level more vivid and clear to see. "√" means having the feature. For example, in the column "Text Length", there is a "√" on the IBT and a "×" on the PBT, meaning text is longer in the IBT than in the PBT. The "≈" means half the feature. "Absence of text while answering" in the IBT is marked by a "≈", meaning while questions are answered, the text is partially absent (see Section 3.2.2). But in the PBT in this column, there is a "×, which means while answering questions in the PBT reading test, text is not available to consult. In the first four columns, one "√" adds one point to the test difficulty of the test version. In the last column, because it is crucial and decisive to test difficulty, question types will be timed by their designing purpose in the level of difficulty in this order: PF<PB< PL< <PI (see p. 18). Questions for PF are timed once, and PB is timed twice, PL timed three times, and PI four times in calculating their contribution to test difficulty.

Test Difficulty		IBT		PBT	
Absence of text while answering		≈	0.5	×	0
Text Length		√	1	×	0
Text Difficulty	**Vocabulary**	√	1	×	0
	Syntax	√	1	×	0
Question Difficulty	**V and S***	√	1	×	0
	Item length	√	1	×	0
Relationship between questions and text	**PF×1**	√	7	√	8
	PB×2	√	4.5	√	1
	PL×3	√	5.5	√	1
	PI×4	√	2.5	×	0
Result		48		13	

Table 16. Test difficulty calculation of the IBT and the PBT sample reading test

The reason for using numbers and calculations here is to make clear whether one reading test is more difficult than the other. The numbers in the "Result" column in table 16 are calculated according to my own method of number calculation, i.e. PL x 3, to demonstrate the relative importance of specific factors, thus they do not show to what extent the IBT sample reading test is more difficult. But numbers in "result" indicate clearly that the IBT sample reading test is more difficult than the PBT one.

Test difficulty has the crucial influence on the results of a test, which are always represented by scores. And test difficulty and scoring method are both crucial factors that are decisive to the reliability of a test. The analysis and discussion on the TOEFL test difficulty indicate that the IBT is more difficult than the PBT in reading test. As Bachman and Palmer claimed (see Section 2.1) that the consistency between the test difficulty and the test result is decisive to test reliability, scoring methods and scoring reports on the TOEFL test will be analyzed as the last step in assessing the TOEFL test reliability.

3.3 Reading test scoring methods and results in TOEFL

As demonstrated above, when a reliable reading test is taken twice by the same test taker, scores he achieves, which show his reading ability, are similar. When a reliable reading test is taken by a certain group of test takers more than once, the average scores they achieve in different times will also be similar. In the case of the TOEFL test, scores of both versions are used for college entrance applications at the same time. Therefore, whether test takers with scores from the two versions can be judged to have the similar reading ability is crucial for the TOEFL test reliability.

When the same person takes both the IBT and the PBT, he will achieve two scores respectively. If reading scores from the IBT and the PBT show the same test taker has different reading ability levels, the TOEFL test reliability should be doubted. If reading scores achieved by the same person in the IBT and the PBT can be interpreted presenting the same reading ability level, the TOEFL test is proved then in this case to be reliable. A large number of scores achieved by various groups of test takers should be analyzed in order to assess the TOEFL test reliability in general.

In the following part in this section, scoring reports of the IBT and the PBT for 2009 will be analyzed comparatively. Scores achieved by test takers of different genders, educational backgrounds and geographic regions will be picked out, and the comparative analysis of the IBT and the PBT reading scores will show whether the same group of test takers' reading abilities are interpreted the same in the two versions of the TOEFL test.

3.3.1 Scoring systems and methods in the IBT and the PBT

Questions in both the IBT and the PBT reading test are multiple-choice questions, which are objectively scored by machine. As it has been demonstrated in former sections that the objectivity of scoring procedures ensures the scoring standard to each test takers is similar and fair, the scoring processes of the TOEFL reading tests in both the IBT and the PBT then can be considered positively to affect the TOEFL scoring reliability. Then the reading test results or scores should be analyzed. Scores are yielded

by scoring process but can be affected by many other factors. The influential factors of scoring reliability are to be discussed later at the end of this section.

In the IBT, "score" and "level" are used for assessing test takers' reading abilities, but in the PBT, only "score" is shown and no reading levels divided by the scores achieved. In the IBT, test takers will not only achieve a score but also get a level of "High", "Intermediate" or "Low". The PBT, on the other hand, only provides score range to test takers. Fortunately, the score comparisons between the IBT and the PBT have been publicized by the official TOEFL website, according to which, the PBT scores can be converted into the IBT scores. Therefore, in the PBT, test takers can also *indirectly* achieve their language ability levels by the converted IBT scores. Due to the different scoring system adopted by the IBT and the PBT, *numbers* of the IBT and the PBT scores are not comparable. Hereafter, reading comprehension *levels* measured by the IBT scores and the PBT scores will be analyzed and compared by specific statistics in order to assess the TOEFL test reliability.

3.3.2 Scoring results of reading tests in the IBT and the PBT

TOEFL assumed that the same test takers will receive the same scores. In converting scores between the IBT and the PBT, the official TOEFL website has published a "scoring comparisons" for the IBT and the PBT whole scores and scores in each part. The following analysis is based on the TOEFL claim and the officially published scoring converting standard. Scoring reports for test takers of different educational backgrounds, different genders, and different nationalities will be used to investigate whether the IBT and the PBT scores achieved by the same group of people are consistent in rated their reading comprehension ability.

To take Africa for example, if the TOEFL test is reliable, scores from the IBT should be consistent with scores from the PBT in showing the average language ability of all African test takers in the same year. Besides geographic region differences and educational background differences, the scoring reports of the IBT and the PBT reading tests will be compared between different genders. If the average scores in IBT and PBT are similar in showing the average language abilities of the same group of test takers, the two versions of the TOEFL test can be then proved to have score reliability.

Three sets of tables will be presented in the following, showing the IBT and the PBT average reading scores of test takers of different educational background, gender, and geographic region. By analyzing the IBT and the PBT average reading scores from the same group of people, the level of score consistency can be explicated.

The first table shows the IBT and the PBT average scores among test takers of different educational levels:

	Mode	Average Score	Percentile Rank
Graduate Students	IBT	21.1	40-45
	PBT	53.2	44-45
Undergraduate Students	IBT	18.5	43-47
	PBT	51.7	42-47
Other Students	IBT	19.7	42-46
	PBT	51.6	56-61
Applicants for Professional License	IBT	21.0	43
	PBT	52.1	43

Table 17. Test takers' average scores and percentile rank in TOEFL—Educational Background (Male*****)

*****In the official TOEFL website, male and female of educational background are issued respectively in presenting the IBT and the PBT scores. Here in the chart above, the statistics of male's are randomly selected out.

From the percentile in table 17, with the exception of the figures in "Other Students", the remaining four groups of test takers of different educational backgrounds have the approximately similar percentiles. To take the undergraduate group, for example, the average score in the IBT is 18.5 and that in the PBT is 51.7. The two scores are of quite the same percentile rank. The close percentiles indicate that there are roughly the same numbers of test takers above 18.5 in the IBT or above 51.7 in the PBT. In 75% of the cases within the same educational groups, the IBT and the PBT percentile ranks are consistent. The other 25% is in the "Other Students" group, in which percentile ranks of

the average scores are greatly different. The possible reason for this phenomenon will be discussed in the end of this section.

Secondly, the table below compares average scores and percentile ranks in the IBT and the PBT between test takers of the two genders.

	Mode	Average Score	Percentile Rank
Male	**IBT**	20.2	42-43
	PBT	52.5	43
Female	**IBT**	19.7	40-42
	PBT	52.2	40

Table 18. Test takers' average scores and percentile rand in TOEFL—Gender

From table 18, the score reliability within each gender group is self-explanatory. Percentile ranks of the IBT and the PBT in the "Male" and "Female" groups are almost the same. It indicates that the two versions of the TOEFL test are consistent in rating test takers reading ability.

Thirdly, in analyzing the IBT and the PBT score consistency, the following two charts present respectively the average reading scores in different geographic regions in the IBT and the PBT.

Region	Number of Participant Countries	Average IBT Score in Reading	Converted Score from the PBT
Africa	37	16.35	16.2
America	25	19.6	19.9
Asia	30	18.3	18.4
Europe	44	21.3	20.6
Middle East & North America	19	15.7	15.1
Pacific Region	2	20	17

Table 19. The average IBT scores in reading test in different geographic regions

Region	Number of Participant Countries	Average PBT Score in Reading	Converted Score from the IBT
Africa	45	51.2	51.35
America	35	53.9	53.6
Asia	28	52.9	52.6
Europe	43	55.6	56.15
Middle East & North America	17	50.1	50.7
Pacific Region	11	52	54.5

Table 20. The average PBT scores in reading test in different geographic regions

The comparisons between each column in table 19 and 20 illustrate that:

- In Africa, the average reading score in the IBT is higher than that in the PBT (see Table 19). This result is not caused by test difficulty difference between the two versions. According to the official TOEFL website, the PBT is mainly used to supplement the IBT in places that do not have internet access (see p. 2). But there are seven more participant countries in the PBT than in the IBT in this region. Test

takers in the seven countries, where internet is not connected and computers are not widely-used, have less materials and sources to learn English or to prepare the TOEFL test than test takers in internet-accessible countries. As a result, the phenomenon in Africa that the mean score of the PBT is lower than the IBT is not a disproof for the TOEFL score reliability. The reason for this fact can be analyzed in further studies of economic influence for language test performance, which is not to be deeply discussed in this essay.

- According to the research in Section 3.2.3 and results achieved in table 16, reading test in the IBT is more difficult than that in the PBT. As a result, reading scores in the IBT should be on average lower than in the PBT. There are two regions, America and Asia (see Table 19, column 2 & 3), among all the six in which the IBT reading scores are on average lower than the PBT ones.
- In Europe and the Middle East, the numbers of participant countries in the IBT and the PBT are almost identical. However, in these two regions, the average scores in the PBT are much lower than in the IBT for 0.55 to 0.6 point. This disparity is larger than in the African case.
- The Pacific region is a special case. Only two countries participate in the IBT, one of which is Australia. Test takers from Australia tend to achieve higher scores. And Australian test takers' scores contribute 50% to the average score of the IBT in this region. In the PBT, there are 11 participant countries, in which Australian test takers' scores contribute less than 10% to the average score. This region is then an exception and not to be considered to affect the TOEFL test score reliability.

After the analysis of the IBT and the PBT scoring reports on test takers of different genders, different education backgrounds and different geographic regions, a specific sample is selected to be analyzed before getting the conclusion of the IBT and the PBT scoring reliability. In the chart that follows, countries in each geographic region that achieve the highest scores (called 'Top Country') in the IBT and the PBT are picked out to be analyzed. Their average scores in the IBT and the PBT are compared. The comparison will show whether the IBT and the PBT testing results are consistent in specific countries and in specific time, the year 2009. This analysis can be seen as a continuous research which is linked to the former analysis in table 19 and 20.

Geographic Region	Mode	Top Country in IBT/PBT	Score	Converted score	Top country'	Score	Converted score
Africa	**IBT**	**Mauritius**	24	58	Zimbabwe	20	54-55
	PBT	**Zimbabwe**	56	21	Mauritius	54	20
America	**IBT**	**Argentina**	24	58	Paraguay	21	56
	PBT	**Paraguay**	54	20	Argentina	56	21
Asia	**IBT**	**Singapore**	25	58-59	Maldives	/	/
	PBT	**Maldives**	66	29	Singapore	62	27
Europe	**IBT**	**Netherlands**	25	58-59	Finland	24	58
	PBT	**Finland**	60	26	Netherlands	59	25-26
Middle East & North America	**IBT**	**Israel**	22	56-57	/	/	/
	PBT	**Israel**	56	21	/	/	/
Pacific Region	**IBT**	**Australia**	20	54-55	/	/	/
	PBT	**Australia**	61	27	/	/	/

Table 21. Top-score countries in the IBT and the PBT in geographic regions

In table 21, when a country achieves the highest scores in the IBT or the PBT, it is called the "Top Country". If a country is a "Top Country" in the IBT but not in the PBT, it is Top Country', in the PBT, in order to be compared its PBT score and converted score with PBT top country's score and its converted IBT score. The comparison can show whether the IBT and the PBT scores indicate consistently test takers' language level. In the "Middle East & North America" and "Pacific Region", Top Country in the IBT and the PBT is the same, so there is no Top Country' in these two regions.

The "Africa" column in table 21 shows the average score of the IBT in Mauritius is 24, the highest in Africa. According to "Score comparisons for reading", the score of 24 can be rated as "High level". The average score of Mauritius in the PBT is 54, converted into the IBT score, its 20, which is rated as "Intermediate level". It indicates

the possibility that a high level reader in the IBT can be rated an intermediate level reader in the PBT. This phenomenon affects the reliability of the TOEFL test negatively.

Based on the statistics in table 21, and using the same method as that used in analyzing Mauritius, the Top country in Africa in the IBT, Argentina and Australia, the IBT Top Countries in "America" and "Pacific Region" are examined. Their PBT scores, when converted into the IBT ones, are not consistent with scores achieved in the IBT in showing test takers' reading ability level. In Australia, converted PBT score degrades the reading level shown by the IBT score. The case of Australia indicates that test takers tend to achieve higher scores in the PBT reading test than in the IBT one. The reason for the phenomenon can be explained by the different test difficulty of the IBT and the PBT. Because the IBT reading test is more difficult, scores in the IBT are lower than in the PBT is reasonable. In the case of Argentina, conversely, the converted PBT score lifts the reading level shown by the IBT score, which is not explicable by test difficulty difference.

With the exception of the above three cases, i.e. Mauritius, Argentina, and Australia, Maldives in this region is special because the IBT is unavailable there. As a result, there is no comparison in the case of Maldives.

In the remaining six Top Countries, the scores in Israel and Netherlands are almost identical: Netherlands test takers' average IBT score is 25, converted into the PBT is 58-59; the PBT average score is 59 and converted IBT score is 25, just the same as the average IBT scores. Similarly, in Israel, the average IBT score and converted IBT score from the PBT scores are consistently in numbers. The TOEFL test then is reliable in Israel and Netherlands because no matter which version test takers take, they show the same reading abilities.

There are still four Top Countries remaining: Zimbabwe, Paraguay, Singapore, and Finland. With the exception of Paraguay, the PBT scores are higher than the IBT ones in the other three countries. According to the analysis of test difficulty in the former sections, the result that the same test takers achieve lower scores in the IBT than in the PBT is predictable and logical. Even though scores in the IBT in these three countries

are lower than in the PBT, the reading abilities rated by both sets of scores are identical. The TOEFL test can then be considered reliable.

The phenomena presented by scores in Paraguay, Argentina and Mauritius are special: since reading test difficulty in the IBT is higher than in the PBT, the fact that the IBT average scores in the reading test are higher than in the PBT reading test are not reasonable. Possible reasons for will be discussion in the end of this section.

3.3.3 Discussion of the TOEFL score reliability

In accordance with the above analysis, the IBT test is more difficult than the PBT. Due to technical reasons, i.e. the formula for calculating the test difficulty in section 3.2.3 (see Table 16, p. 37) is created by me according to relevant theories, the analysis of the IBT and the PBT test difficulty comparison is from qualitative, not quantitative perspective. As a result, it is impossible to know how much lower the IBT scores are than the PBT. Theoretically speaking, however, it is possible to show and offer reasons for why the IBT scores are lower, or at least, no higher than the PBT scores. Based on the comparison between the IBT and the PBT average scores in different groups of people, especially in top-score countries, the implications of such differences can be identified as follows:

- Questions in both versions are all multiple-choice, and manually scored. The objectiveness of scoring method affects score reliability positively.
- According to table 18, the IBT and the PBT test takers' average scores have a relatively similar percentile rank, which indicates test takers' performances in the two versions are quite steady.
- According the table 17, with the exception of "Other Students" item, 75% of all the test takers of different educational backgrounds achieve relatively similar scores and are rated to have similar reading ability levels by both versions.
- According to table 19 and 20, the TOEFL scores are reliable in 40% of all the geographic regions (America and Asia); 20% (Africa) of all the cases is influenced by economic reasons; in 40% of the regions, namely Europe and Middle East & North America, the TOEFL test scores are not explanatory as

reliable ones in taking the test difficulty difference between the IBT and the PBT reading tests into consideration.

- According to table 21, represented by the Top Country scores in the IBT and the PBT, the TOEFL scores are reliable in 60% of the geographical regions. Though there are test difficulty differences, scoring system differences and delivering mode differences, the TOEFL scores are consistent irrespective of the type, i.e. internet- or paper-based of test taken. Take Australia for instance, though the average score in the IBT is much lower than that in the PBT, TOEFL can still be considered of score reliability because the IBT is more difficult than the PBT in reading tests.
- The consistency of the TOEFL score reliability can not be assessed in some regions, such as Maldives because the IBT is not available there. This phenomenon indicates that in some areas the IBT and the PBT are incomparable because of the technical limitation. It implies that the scientific and technical development is also an important factor that influences the TOEFL test results.
- For The remaining 30% of the regions, average scores in the IBT are higher than in the PBT. Additionally, in 2/3 of them, scores in the IBT and the PBT rate test takers' reading abilities into different levels. In some situations, the inconsistency between two versions of the TOEFL test can cause test takers to fail in college entrance applications.
- In groups of 'Middle East & North America', and 'Europe', scores in the PBT and in the IBT are not consistent. Test takers from these two regions get much higher scores in the IBT than in the PBT, and this fact is very hard to explain based on the present research done in this essay.
- In 'Paraguay', 'Argentina', and 'Mauritius', the average scores in the IBT are higher than in the PBT. The results affect negatively the TOEFL score reliability.

The following section will discuss in detail the TOEFL test reliability in terms of reading, in relation to the test settings, test difficulties and scoring systems in the two versions of test, the IBT and the PBT.

3.3.4 Discussion of the TOEFL test reliability

Firstly, test setting analysis shows that the IBT and the PBT are varied in organizations and delivering modes. But these differences in the IBT and the PBT settings are external features and then not decisive factors for the consistency of the results the TOEFL test. Additionally, there are both positively and negatively influential factors in the IBT and the PBT test setting. For these reasons above, test settings of the IBT and the PBT are considered not to affect the TOEFL test result consistency.

Secondly, the objective scoring methods in the IBT and the PBT ensure that scores in both of the tests are reliable, with no examiners' personal influence or confusion in scoring criteria.

Thirdly, test difficulty analysis of the IBT and the PBT reading tests indicates that at least from the analysis of sample reading tests, the IBT reading test is more difficult. And the test difficulty difference between the IBT and the PBT can influence the test results of the two versions: it is easy to explain and understand if the IBT reading scores are lower than that in the PBT. It is also possible and acceptable in the cases that the IBT reading scores are quite similar to the PBT ones, because internet facilitates English learning greatly and test takers in the IBT available countries and areas can be of relatively higher English level.

Generally speaking, the TOEFL test is reliable with its two versions. However, there are still some phenomena that are very difficult to explain with regard to test difficulties, i.e. in some geographic regions, the average IBT scores are much higher than the PBT ones. The phenomenon is not explainable until economic development and other social factors are taken into consideration. Test difficulty then can only be a theoretical decisive factor for test result. Test results actually can be influenced greatly by many other factors. Since socio-economic impact is not what the thesis is about, it is not to be enlarged on here.

4. Conclusion

To assess the reliability of the TOEFL test involves tremendous amount of work. Fortunately, a series of TOEFL research has been done by TOEFL researchers all over the world. ETS, as the organizer of the TOEFL test, has edited and electronically published a series of academic monograph on TOEFL research and there have been numerous thesis and documents in about the TOEFL test, including the reliability of the IBT and that of the PBT.

Mainly based on Bachman and Palmer's definition on test reliability and Hughes' theory on influential factors for reliability, this essay focuses on reading test specifically, comparing the PBT with the IBT in test settings, test difficulties and scoring methods. This essay studies on the reliability of the TOEFL test as a whole, with two different delivering versions, and assesses whether the IBT and the PBT are consistent in presenting test takers' English reading comprehension ability and whether the testing processes and results are coherent to the designed purpose of the TOEFL test, which is to test English abilities in academic settings or English college classrooms.

The reliability of the TOEFL test is not obviously affected by either the IBT or the PBT test settings because both of the test settings have positive and negative effects on test results: the IBT, on the one hand, is more convenient and flexible for test takers to choose test time and site; on the other hand, the paper and pencil delivering mode of the PBT, as a traditional one, is more familiar to test takers. Test description in the IBT is more explicit than in the PBT, but former test papers of the PBT, which contain test descriptions specifically, can be used for free on the internet, while the IBT former test papers are not provided for free. Thus test takers can know more clearly about what texts and questions look like before taking the PBT than before taking the IBT. Conclusively, there are test setting differences between the two versions, but they do not cause much difference in the results of the IBT and the PBT.

The comparative analysis of the IBT and the PBT reading tests difficulty is based on the text and questions of sample reading tests of the two versions. And the IBT reading

test, at least represented by the sample test, is found to be more difficult than the PBT one. As the internal factor of a test, test difficulty is the decisive factor for test results. But different scoring methods and scoring systems within the same test can also affect the test results greatly. In the case of the TOEFL test, when scores are converted between standard ones and raw ones, the level they mark can be changed.

From the comparison within TOEFL between its two versions, and the discussions, it is therefore concluded that the TOEFL test in terms of reading has not yet proved or disproved to be reliable by the analysis done in this essay: one the one hand, there are many cases showing that even though the IBT and the PBT are very different in test settings, test difficulties and scoring systems, the TOEFL test results are reliable; On the other hand, in many other cases, test takers achieve much higher scores in the more difficult version than in the relatively easier one and test takers' reading comprehension abilities are evaluated verily in the IBT and the PBT.

There can be technical reasons for the fact that in some cases the TOEFL test tends to be unreliable: since this essay uses the sample reading test papers for the IBT and the PBT in analyzing test difficulty of the TOEFL reading comprehension tests, and in comparing reading scores in the IBT and the PBT, this essay uses scoring reports for 2009. Sample test papers technically represent the test difficulties of real test papers; however, there can always be exceptions. Without analyzing the real test papers for 2009, it cannot be definitely taken for granted that the IBT reading test papers are all more difficult than the PBT ones.

With the exception of the possible reason mentioned above, social-economical factors in some regions can affect greatly or even decisively. Further research should be carried out to investigate the influence of social-economical factors on the TOEFL test results. But this area is not the focus of this essay.

This essay aims to assess the TOEFL test reliability in terms of reading by mainly focusing on test difficulty and test results. Since the TOEFL results affect people and colleges so crucially, further study on the TOEFL test is called for: the IBT and the PBT tests can be done by a group of selected non-native English people, with various social-economical backgrounds. When their scores in the two versions are collected and

compared, whether the TOEFL test results are consistent in the IBT and the PBT can be more directly and clearly investigated.

List of References

Primary Materials

Bulletin for Internet-based Testing (IBT) 2009-2010 Information and Registration. (2009). [online] Available from World Wide Web: <http://www.ets.org/toefl.html> [Accessed April, 2010].

Bulletin for Paper-based Testing (PBT) 2009-2010 Information and Registration. (2009). [online] Available from World Wide Web: <http://www.ets.org/toefl.html> [Accessed April, 2010].

Sample reading test of TOEFL IBT. [online] Available from World Wide Web: <http://www.vanandy.com/toefldownload.html> [Accessed May, 2010].

Sample reading test of TOEFL PBT. [online] Available from World Wide Web: <http://www.vanandy.com/toefldownload.html> [Accessed May, 2010].

TOEFL Performance Feedback for Test Takers. Assessed 13th May 2010. <http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Perf_Feedback.pdf>

Secondary Materials

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Hudson, T 1996, *Assessing Second Language Academic Reading from a Communicative Competence Perspective: Relevance for TOEFL 2000:* <http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect_ets.org_toefl> [Accessed 20th April, 2010].

Hughes, A. (2001). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

IBT Descriptions. [online] Available from World Wide Web: <http://www.interface.edu.pk/tests/toefl/toefl-test-formats.asp> [Accessed 6th May, 2010].

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (1999). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Report No. MS-16). Princeton, NJ: Educational Testing Service

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and practice in assessing second language reading*. Cambridge, UK: Cambridge University Press.

Lee, Y.-W., and Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes.* ETS Research Report (RR-05-14).Princeton, NJ: ETS. <http://infoport/sites/rrpts/publications/RR-05-14.pdf> [Accessed 5th May, 2010].

Megginson, D 2009, *Sentence Structure Identification Samples.* Accessed 11th May, 2010, <http://www.writingcentre.uottawa.ca/hypergrammar/rvsntstr.html>

Munby, J. L. (1968). *Reading and Thinking: Training the Intensive Reading Skills*. Cambridge, UK: Cambridge University Press.

PBT Descriptions. Accessed 3ed May 2010 <http://www.interface.edu.pk/tests/toefl/toefl-test-formats.asp>

Ronald, D. E. (1964). *Milton's Grammar.* The Hague: Mouton.

Sentence Types in English. Assessed 5th May 2010, <http://esl.fis.edu/learners/advice/syntax.htm>

TOEFL IBT and PBT Description Comparison. Assessed 5th May 2010, <http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=d35ed898c84f4010VgnVCM10000022f95190RCRD>

TOEFL Score Reliability. Assessed 8th May 2010, <http://www.chinazzwb.com/news-abroad/12241495.html>

What does a score mean? 2009. Assessed 13th May 2010, <http://personality-project.org/perproj/reliability.html>